Properties of Estimators (2)

STA 211: The Mathematics of Regression

Yue Jiang

February 21, 2023

The following material was used by Yue Jiang during a live lecture.

Without the accompanying oral comments, the text is incomplete as a record of the presentation.

An important disclaimer

This is not a mathematical statistics class. There are semester-long (and multiple semester-long) courses on probability, and so what we cover in just two lectures scarcely touches on even the basics.

However, familiarity with some of these concepts, such as probability distributions and certain aspects of them (e.g., expectation, variance, etc.) are needed to more fully grasp linear models. As such, we will be presenting a very abridged treatment of some of the fundamentals needed to proceed.

Convergence of sequences

Thinking back to calculus, we've learned that a sequence of real numbers $X_1, X_2, X_3, \cdots$ *converges* to a limit X if we can find, for every $\epsilon > 0$, some natural number N such that for every $n \geq N$, $|X_n - X| < \epsilon$.

But what about a sequence of *random variables*? What might it mean for a sequence of random variables "to converge"?

Some difficulties

Suppose $X_1, X_2, X_3, \cdots$ is a random sequence, where each X_i is i.i.d. from a $N(0,1)$ distribution. Can we say that X_n "converges" to a random variable $X \sim N(0,1)$?

$P(X_n = X) = 0$ for every n. Is this a problem?

Some difficulties

How about $X_1, X_2, X_3, \cdots$, where $X_i \sim N(0, 1/n)$? Can we say that X_n "converges" to 0 somehow?

$P(X_n = 0) = 0$ for all n. Is this a problem?

Convergence of random variables

What does it mean for a sequence of random variables "to converge"? (kind of a trick question for now)

Chebyshev's Inequality

Let's take a quick break from notions of convergence.

Chebyshev's Inequality provides a statement about "how far we can be" from the mean for any probability distribution, as long as we know something about their expectations and variances. In particular, if a random variable X has finite $E(X) = \mu$ and $Var(X) = \sigma^2$, then:

$$P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$$

- Take a moment to look at the statement above. What does it mean "in plain English"?

Example: the sample mean

Let's consider the sample mean $\bar{X}$ of n i.i.d. random variables $X_1, \cdots, X_n$, where X_i has finite expectation and variance $E(X) = \mu$ and $Var(X) = \sigma^2$.

- What are $E(\bar{X})$ and $Var(\bar{X})$?
- Can you provide a bound on $P(|\bar{X} - E(\bar{X})| \geq \epsilon)$?

Example: the sample mean

$$
\begin{aligned}
E(\bar{X}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n} E\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \\
&= \frac{1}{n} n E(X) \\
&= E(X) = \mu.
\end{aligned}
$$

Example: the sample mean

$$
\begin{aligned}
Var(\bar{X}) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n} Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j)\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \\
&= \frac{1}{n^2} n Var(X) \\
&= \sigma^2/n
\end{aligned}
$$

Example: the sample mean

Thus, Chebyshev's inequality gives us

$$P(|\bar{X} - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$

- What happens as the sample size n goes to infinity?

A consequence of Chebyshev's inequality

For i.i.d. random variables $X_1, \cdots, X_n$ with finite expectation μ and variance σ^2,

$$\lim_{n \to \infty} P(|\bar{X} - \mu| \geq \epsilon) = 0.$$

Technically, the statement is also true even if we don't require finite variance - this statement is known as the **weak law of large numbers**.

- What does the WLLN mean, intuitively?

The weak law of large numbers

For i.i.d. random variables $X_1, \cdots, X_n$ with finite expectation μ

$$\lim_{n \to \infty} P(|\bar{X} - \mu| \geq \epsilon) = 0.$$

As n gets larger and larger, the *sample mean* will be within any ϵ of μ with a "high probability." As n goes to infinity, the probability that the sample mean is farther than ϵ from μ goes to 0.

Convergence in probability

A sequence of random variables $X_1, X_2, X_3, \cdots$ is said to converge in probability to a *random variable* X if

$$\lim_{n \to \infty} P(|X_n - X| \geq \epsilon) = 0,$$

for all $\epsilon > 0$.

An estimator that is **consistent** is one that converges in probability to the parameter of interest. For instance, we saw that for the example on the previous slide, the sample mean was a consistent estimator for the population mean.

- What is the "sequence" of random variables here?

Consistency and unbiasedness

- Is a consistent estimator always unbiased?
- Is an unbiased estimator always consistent?

Consistency and unbiasedness

Consider a random sequence of estimators of the population mean (where $E(X) = \mu$, and let's suppose $Var(X) = \sigma^2 < \infty$ for simplicity), given by $X_n = \bar{X} + \frac{1}{n}$. This is indeed a consistent estimator for the mean μ, since

$$lim_{n\to\infty} P\left(\left|\bar{X} + \frac{1}{n} - \mu\right| \geq \epsilon\right) = 0,$$

even though it is biased (the expectation is $\mu + \frac{1}{n}$, which is not equal to μ).

Consistency and unbiasedness

Consider a the estimator of the population variance for an $N(\mu, \sigma^2)$ distribution given by $\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$. As we've seen on HW 6, this is a biased estimator for σ^2. However, it is **consistent** (homework).

Consistency and unbiasedness

Consider a random sequence of estimators of the population mean for an $N(\mu, \sigma^2)$ distribution, given by $X_n = X_1$. This is unbiased, but not consistent for μ (Why? What is the variance of this estimator?)

Consistency and unbiasedness

(See board).

A sufficient condition for consistency

With that said, if you have a consistent estimator whose variance goes to zero asymptotically, then this estimator is also consistent (for the parameter of interest). This is a pretty common strategy for demonstrating consistency of estimators.

As an aside, anyone know a one-line reason for why this is the case? (if you do, you really shouldn't be in this class!)

Consistency of the OLS estimator

We've previously established that the OLS estimator is unbiased for $\boldsymbol{\beta}$. Is it also a consistent estimator?

Note - we'll be generalizing the definition of convergence in probability to the multivariate case for random vectors, but this isn't too bad of a problem. Choose your favorite vector norm; we won't go through the details, but it's basically the same definition just using the distance between vectors.

Consistency of the OLS estimator

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\
&= \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\epsilon} \\
&= \boldsymbol{\beta} + \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\epsilon_i\right),
\end{aligned}$$

where $\mathbf{x}_i$ represents the i^{th} row of the design matrix $\mathbf{X}$.

Consistency of the OLS estimator

Then under some assumptions (which ones?), the WLLN says that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T \rightarrow_p E(\mathbf{x}_i\mathbf{x}_i^T),$$

which is some non-singular matrix, and

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\epsilon_i \rightarrow_p E(\mathbf{x}_i\epsilon_i) = \mathbf{0}.$$

Consistency of the OLS estimator

Thus,

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T\right)^{-1}}_{\to_p E(\mathbf{x}_i\mathbf{x}_i^T)^{-1}} \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\epsilon_i\right)}_{\to_p \mathbf{0}}$$

$$\to_p \boldsymbol{\beta}$$

demonstrating consistency. Under weak assumptions, the OLS estimator is both unbiased *and* consistent for $\boldsymbol{\beta}$.

(there were a few details we skipped, mostly the use of Slutsky's Theorem and the continuous mapping theorem and "why/what" "converging" in probability to a matrix means, but this will be covered in more detail in an actual math stats course - the intuition and broad details are what's important here).

A different type of convergence

We just explored **convergence in probability**, which described a notion of convergence where random variables "converge" if there is a low probability of them being "very far" from each other.

What about some notion of distance based on whether their distribution functions are "close" to each other? Can we derive some notion of "convergence" in this sense?

Convergence in distribution

Consider a sequence $X_1, X_2, X_3, \cdots$ of random variables, and let F_{X_n} be the distribution function of this sequence.

The sequence $X_1, X_2, X_3, \cdots$ is said to be **convergent in distribution** to a random variable X if

$$\lim_{n \to \infty} F_{X_n}(x) = F_X(x)$$

at all continuity points of F_X. We commonly call X the "asymptotic distribution" or "limiting distribution" of the sequence of random variables $\{X_n\}$.

An example

Consider the random sequence $X_1, X_2, X_3, \cdots \overset{i.i.d.}{\sim} Unif(0,1)$. Let $X^{(n)}$ be the maximum. Note the following for any $\epsilon > 0$:

$$\begin{aligned} P(X^{(n)} \leq 1-\epsilon) &= P(X_1 \leq 1-\epsilon, X_2 \leq 1-\epsilon, \cdots) \\ &= (1-\epsilon)^n. \end{aligned}$$

Setting $\epsilon = t/n$:

$$\begin{aligned} P(X^{(n)} \leq 1-t/n) &= (1-t/n)^n \\ P(n(1-X^{(n)}) \geq t) &= (1-t/n)^n \\ P(n(1-X^{(n)}) \leq t) &= 1-(1-t/n)^n \\ &\to 1-e^{-t}, \end{aligned}$$

which is the distribution function of $Exp(1)$. So the random variable $n(1-X^{(n)}) \to_d Exp(1)$.

Another example

Suppose you have a random sequence of i.i.d. $X_1, X_2, X_3, \cdots$ all drawn from a population with finite mean μ and variance σ^2. Then we have

$$\sqrt{n}(\bar{X}_n - \mu) \to_d N(0, \sigma^2)$$

- What is this result?
- What is the importance of scaling by $\sqrt{n}$?

Asymptotic normality of OLS estimator

Consider the term $\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\epsilon_i$. This "looks like" a sample mean of some sort, and we might expect some CLT-like result to hold. We additionally know from before that the expectation of this quantity is $\mathbf{0}$.

In fact, the CLT (well, the multivariate case, but don't worry too much) precisely tells us that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\epsilon_i - \mathbf{0}\right) \to_d N(\mathbf{0}, \Sigma),$$

where $\Sigma = \sigma^2 E(\mathbf{x}_i\mathbf{x}_i^T)$.

Asymptotic normality of OLS estimator

Then in fact, appropriately scaling the $\widehat{\boldsymbol{\beta}}$ term, we'll get

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) = \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T\right)^{-1}}_{\to_p E(\mathbf{x}_i\mathbf{x}_i^T)^{-1}} \underbrace{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\epsilon_i\right)}_{\to_d N(\mathbf{0},\sigma^2 E(\mathbf{x}_i\mathbf{x}_i^T))}$$
$$\to_d N(\mathbf{0}, \sigma^2 E(\mathbf{x}_i\mathbf{x}_i^T)^{-1}).$$

And so we have that the OLS estimator, appropriately scaled, is also **asymptotically normal**.

(we skipped some more details again, primarily around Slutsky's Theorem and the Cramer-Wold Theorem. Again, this will be covered in more detail in an actual math stats course - the intuition and broad details are what's important here).

Homework 7: Due Mar. 7

1. Consider again the estimate of the population variance for an $N(\mu, \sigma^2)$ distribution given by $\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Demonstrate that it is a consistent estimator. You may use the fact that if you have an i.i.d. sample from a $N(\mu, \sigma^2)$ distribution, then the variance of the sample variance, $Var(s^2) = \frac{2\sigma^4}{n-1}$. For this problem you may assume $\sigma^4 < \infty$.
2. Let X be 1 with probability 0.5 and 0 with probability 0.5 (that is, $Bern(0.5)$). Let $X_n = X$, and $Y = 1 - X$. Show that $X_n \to_d Y$ but that $X_n \not\to_p Y$ (hint: consider how "far apart" X_n and Y are, then use the definition of convergence in probability for an appropriate ϵ).
3. Suppose $X_1, \cdots, X_n$ are an i.i.d. sample from a distribution with density $f_X(x) = \frac{\lambda x + 1}{2}$ for $x \in (-1, 1)$ and $\lambda \in (-1, 1)$. Consider the estimator $3\bar{X}$. Is it a biased estimator for λ? Is it a consistent estimator for λ? Show your work and explain.